



# Gaze-assisted automatic captioning of fetal ultrasound videos using three-way multi-modal deep neural networks

Mohammad Alsharid<sup>a,c,\*</sup>, Yifan Cai<sup>a,1</sup>, Harshita Sharma<sup>a</sup>, Lior Drukker<sup>b,d</sup>,  
Aris T. Papageorghiou<sup>b</sup>, J. Alison Noble<sup>a</sup>

<sup>a</sup> Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, OX3 7DQ, United Kingdom

<sup>b</sup> Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, OX3 9DU, United Kingdom

<sup>c</sup> Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, 127788, United Arab Emirates

<sup>d</sup> Women's Ultrasound, Department of Obstetrics and Gynecology, Beilinson Medical Center, Sackler Faculty of Medicine, Tel-Aviv University, Tel Aviv, Ramat Aviv 69978, Israel

## ARTICLE INFO

### MSC:

92C55

68T45

68T50

97U80

### Keywords:

Video captioning

Gaze tracking

Fetal ultrasound

Audio-visual

Multi-modal

## ABSTRACT

In this work, we present a novel gaze-assisted natural language processing (NLP)-based video captioning model to describe routine second-trimester fetal ultrasound scan videos in a vocabulary of spoken sonography. The primary novelty of our multi-modal approach is that the learned video captioning model is built using a combination of ultrasound video, tracked gaze and textual transcriptions from speech recordings. The textual captions that describe the spatio-temporal scan video content are learnt from sonographer speech recordings. The generation of captions is assisted by sonographer gaze-tracking information reflecting their visual attention while performing live-imaging and interpreting a frozen image. To evaluate the effect of adding, or withholding, different forms of gaze on the video model, we compare spatio-temporal deep networks trained using three multi-modal configurations, namely: (1) a gaze-less neural network with only text and video as input, (2) a neural network additionally using real sonographer gaze in the form of attention maps, and (3) a neural network using automatically-predicted gaze in the form of saliency maps instead. We assess algorithm performance through established general text-based metrics (BLEU, ROUGE-L, F1 score), a domain-specific metric (ARS), and metrics that consider the richness and efficiency of the generated captions with respect to the scan video. Results show that the proposed gaze-assisted models can generate richer and more diverse captions for clinical fetal ultrasound scan videos than those without gaze at the expense of the perceived sentence structure. The results also show that the generated captions are similar to sonographer speech in terms of discussing the visual content and the scanning actions performed.

## 1. Introduction

Image captioning is a task at the intersection of Natural Language Processing (NLP) and Computer Vision (CV) with the aim of automatically generating descriptions for visual cues in images (Bernardi et al., 2016; Allaouzi et al., 2018; Alsharid et al., 2019). Captioning stands out from other NLP-based tasks as the input itself is not in the form of natural language but a visual stimulus (Allaouzi et al., 2018; Kougia et al., 2019). Video captioning, the subject of this paper, extends the concept of image captioning by exploiting rich, spatio-temporal information in the video clip of interest (Pan et al., 2016b). In medical imaging, video capture and analysis is not widely employed in a clinical setting; however, capturing medical imaging video data,

and specifically ultrasound imaging, has allowed for different kinds of analyses to occur (Drukker et al., 2020a,b). Video captioning can be useful in analysing fetal ultrasound content because of the high temporal resolution of ultrasound. Our interest in this paper is to investigate automatic ultrasound video captioning. In particular, in this paper, the objective is to build a video captioning method to describe routine second-trimester fetal ultrasound scans in the sonographer spoken vocabulary. A novelty of our work is that we pose this as a *multi-modal* problem, using ultrasound video, audio, and gaze as input. We propose an original multi-modal deep learning framework to describe ultrasound (US) scan videos assisted by sonographer gaze information. To the best of our knowledge, this is also the first attempt

\* Corresponding author.

E-mail address: [mohammad.alsharid@eng.ox.ac.uk](mailto:mohammad.alsharid@eng.ox.ac.uk) (M. Alsharid).

<sup>1</sup> Equal contribution.

<sup>2</sup> Work done during PhD.

to perform automatic video captioning using sonographer descriptions derived from their speech recordings, that is, a captioning method developed for live scan videos rather than images in fetal ultrasound scans. In the latter case when captioning images, the focus is only on interpretation. The spatio-temporal video captioning model learns from a comprehensive vocabulary of sonographer speech while describing US scans in terms of the anatomical content and the scanning actions performed, including search, fine-tuning and interpretation.

Medical images can be challenging to describe for the layperson who does not possess the sufficient expert knowledge needed to do so. In a situation, where an expert sonographer cannot be present to perform and describe the scan, such as in remote areas, a tool of this nature may be of assistance. Such computational methodology, fully developed and validated, might in the future provide an educational tool for trainees, support occasional users of US, or assist in conveying information to a patient about a scan in an intuitive way. In essence, such a tool can potentially have a wide appeal and prove to be useful to a number of different users. In this paper, we build a deep learning model from audio–video data that can potentially serve as the foundational building block of a prototype of such a tool.

As with medical images, the analysis of medical video clips, particularly ultrasound, have associated challenges, including a scarcity of labelled data and medically qualified annotators, anatomical class imbalance, and, fortunately, an over-abundance of healthy cases (Allaouzi et al., 2018; Alsharid et al., 2019; Kougia et al., 2019). In our work, we enrich our models, making them more accurate, by leveraging data from another modality, eye tracking data. There is additional novelty in leveraging sonographer gaze-tracking information to aid the video captioning process, as the model generates text that reflects the sonographer visual attention during the acquisition and spoken description of the scan video. We hypothesise that the additional information prevalent in the eye tracking data will help build better performing captioning models.

Previously, a number of image captioning models originally developed for natural images have been studied in the medical imaging domain, such as X-ray images (Xiong et al., 2019) and US images (Alsharid et al., 2019). In some work, captioning has been proposed to generate automated reports for preliminary diagnosis (Allaouzi et al., 2018; Kougia et al., 2019); however, in obstetric ultrasound, previous work has argued the potential value not only for diagnostic reporting but also communicating findings to a non-expert unfamiliar with interpreting ultrasound scans (Alsharid et al., 2019). Alsharid et al. (2019) also reduces the need for directly labelled data from the sonographers by introducing an approach to retrospectively collect speech recordings and converting them to textual descriptions. In the current work, we used a different method to collect speech recordings on-the-fly during real-world scans. These speech recordings are transcribed into time-stamped words, which are then combined to serve as captions for the clips.

Visual attention, as described in Cai (2019), refers to the actual human attention in the form of their gaze behaviour when viewing video clips or images. The motivation behind using gaze for visual attention is the hypothesis that the gaze information could help in training better deep learning models. This hypothesis was investigated and proven by Cai (2019) for the task of standard plane detection.

### 1.1. Clinical motivation

As part of routine care, pregnant women are offered a detailed fetal anomaly ultrasound scan at approximately 20 weeks of gestation to identify any fetal malformations and to assess the current anatomical development of the fetus and its growth. Captioning second trimester videos is interesting because of multiple fetal anatomical development parameters assessed and the anomaly checks performed. These required checks are regulated in the UK by the National Health Service (NHS) as part of the Fetal Anomaly Screening Programme (FASP) (Kirwan,

2010). However, in which order these checks are made, how much time to dedicate to each check, and whether or not to repeat a check is up to the sonographer performing the scan. The structures to be viewed include the spine, the abdomen, the head, and the heart. Structures could be viewed in different planes (such as sagittal) and imaging modes (such as colour Doppler) (Sharma et al., 2021). The way the probe is used and moved can also vary in second trimester scans. Medical images can be challenging to describe for the layperson who does not possess the sufficient expert knowledge needed to do so. A tool that can be of assistance in such scenarios can potentially have a wide appeal and prove to be useful to a number of different users. In this paper, we build machine learning models from audio–video data that can potentially serve as the foundational building blocks of a prototype of such a tool. However, we must keep in consideration the fact that datasets used in medical image analysis come from the real world, and such datasets are often small-sized, in terms of number of data samples, and imbalanced. Table 1 shows how our dataset compares with established datasets used for benchmarking in video description related tasks. Those datasets are built through crowd-sourcing or by acquiring existing videos made available on YouTube. For that reason, it is essential that we carefully consider how to prepare the data and how to train a model in such a way that we mitigate the difficulties associated with using real world data.

The technical idea explored in this paper is to investigate the development of a framework for interpreting ultrasound (US) video automatically and conveying this interpretation in (English) text form.

The most obvious possible application for text generation in a medical imaging context would be report generation or diagnostics, and there is recent work that has been done for that in the case of diagnostics (Zeng et al., 2018, 2019). In our work, however, the raw and temporal nature of our collected data allows us instead to look at more interesting aspects, such as describing the scanning activity, and for more nuanced applications that are only possible when in possession of the sonographers' audio recordings that accompany the fetal ultrasound scans and the transcribed textual form of those audio recordings combined with the corresponding visual content.

The aim of our work is to learn joint image–text representations to describe ultrasound images with rich vocabulary consisting of nouns, verbs, and adjectives. The resulting deep learning models may be useful, in the future, if embedded within systems developed to aid in interpreting the contents of frames and clips from ultrasound scan videos.

### 1.2. Related works

#### 1.2.1. Image and video captioning

Image captioning methods can be broadly separated into three categories (Bernardi et al., 2016; Tanti et al., 2017; Allaouzi et al., 2018; Alsharid et al., 2019); template-based, retrieval-based (Ordonez et al., 2011) and generation-based (You et al., 2016; Elliott and Keller, 2013). Template-based methods typically have a fixed template sentence with blanks which the model fills in with the appropriate words. This approach can result in a lack of flexibility in the potential output as produced captions end up being very similar in terms of structure and grammar. Retrieval-based methods rely on finding visually similar images from a training set to the image in question and then associating the caption of the most similar image(s). Generation-based methods typically rely on deep learning models consisting of an encoding Convolutional Neural Network (CNN) to describe an image, and a textual caption is generated by learning joint image–text embeddings, with Recurrent Neural Networks (RNNs) (Vinyals et al., 2015; Tanti et al., 2017; Alsharid et al., 2019) or transformers (Zhu et al., 2018; Xiong et al., 2019) serving as language models. In the current paper, we consider the generation-based methods for videos, describing their spatio-temporal visual content with words from the expert vocabulary of medical professionals.

**Table 1**

Comparing our realistically sized medical dataset with established datasets used for traditional benchmarking and (pre-)training purposes for tasks involving natural visual scenes.

Dataset	Videos	Duration	Source	Year
MSR-VTT (Xu et al., 2016)	7180	40 h	YouTube	2016
YouCook2 (Zhou et al., 2020)	2000	176 h	YouTube	2018
EPIC-KITCHENS (Damen et al., 2018)	432	55 h	Crowd-Sourcing	2018
HowTo100M (Miech et al., 2019)	1,221,000	134,472 h	YouTube	2019
(Caption-Annotated) PULSE (Ours)	10	5.333 h	Hospital	2019

Many video captioning works have been inspired by work done for machine translation (Wu et al., 2017; Pan et al., 2016a; Yu et al., 2016; Xu et al., 2016), where spatio-temporal visual information is extracted through CNNs and an LSTM-RNN is used to generate the captions given visual information as a prior. The visual information from the videos is being ‘translated’ into text form. Other more traditional work that attempt to caption videos make use of templates with hand-crafted grammatical rules (Kojima et al., 2002; Rohrbach et al., 2013, 2014; Guadarrama et al., 2013; Xu et al., 2015b). The templates have special positions for different parts of a sentence, such as the sentence’s subject, verb, and object. Some of these works then rely on detecting visual concepts in the video to then fill in the template caption with the appropriate words.

In our previous work (Alsharid et al., 2019), we built an image captioning model for still fetal ultrasound images. In Alsharid et al. (2020, 2021), we went further by introducing a curriculum learning based approach to train such image captioning models. We also proposed a captioning framework where an image is first classified before one of many image captioning models are initiated to then caption the image. Each image captioning model is associated with one anatomical structure. The current paper is different in two primary ways. First, it tackles the challenge of captioning fetal ultrasound video clips rather than still fetal ultrasound images. Second, it explores the benefit of introducing and making use of eye tracking information in training. At test time, depending on the model configuration, the gaze-assisted captioning model will expect attention maps merged with the video clips’ frames as input. This map is either prepared from real gaze points or is the output of a saliency prediction model. More details on both configurations and how to prepare their gaze input is discussed in the Methods section.

### 1.2.2. Video description and activity recognition

Video captioning models should describe the objects and what actions, if any, are occurring to or are caused by these objects in a video clip. They differ from video tagging (Siersdorfer et al., 2009; Yao et al., 2013) in that video tagging only generates a set of keywords that are relevant to the video while video captioning aims to generate a sentence or a phrase describing the video clip with the generated caption likely containing many of the same keywords.

Under video description and activity recognition, there are primarily two deep learning approaches to learn spatio-temporal visual information (Wu et al., 2017). The first approach employs 3D convolutional neural networks (CNNs) to accommodate temporal information in the third dimension. However, a significant amount of annotated data is required to train a 3D CNN (Wu et al., 2017), and this is not typically available for medical imaging problems. The second approach employs standard 2D CNNs and learns temporal dependencies between the frames of a video clip by passing the image (frame) feature vectors extracted by the 2D CNNs to a recurrent neural network (RNN). Due to the flexibility of the latter approach in transfer learning tasks where limited training data is available, we have selected this approach for our work in which we use a convolutional LSTM-based RNN architecture to capture and extract the spatio-temporal visual information of a video clip, effectively placing the visual modality under the purview of the convolutional LSTM-based RNN. In addition, 3D CNNs and Long-term Recurrent Convolutional Networks (LRCNs) can be used in conjunction with one another, as was done in Sharma et al. (2021).

### 1.2.3. Incorporating visual attention

Sugano and Bulling (2016) proposed to incorporate gaze tracking information in automatic captioning. We discuss this paper in detail specifically, since it is the primary paper that uses gaze tracking information in the captioning task. Each image was split into  $L$  regions. Each image region has its own feature vector. There is a corresponding scalar value representing the fixation for each region. This value reflects where in the image (which regions) is the gaze most prominent. The gaze was in the form of a histogram of fixation durations or set of fixations. This set of fixations reflects the state of the gaze in that image region. The reason that the gaze is used in the form of a histogram is to allow the neural image captioning model that incorporates visual attention to focus attention differently to each region based on the gaze state in its corresponding location (index) in the fixation set (Sugano and Bulling, 2016). The image was separated into regions with each region having a corresponding fixation from the histogram. Incorporating the gaze information involved the value of a fixation being multiplied with a feature representation of its corresponding region (Sugano and Bulling, 2016). To encourage thorough exploration of the visual scene, Sugano and Bulling (2016) used both the human visual attention and computed soft attention (Xu et al., 2015a) to filter the visual features extracted by the encoding CNN. In contrast, in our case, the ground truth visual attention is computed as a 2D attention map around Cartesian coordinates of gaze points (Cai et al., 2018b). Different from Sugano and Bulling (2016), the current paper modifies the attention filtering mechanism (Cai et al., 2018b, 2020) by implementing a residual operation (He et al., 2016), which is more efficient (Cai et al., 2018b) than calculating soft attention. Das et al. (2017) prepared a dataset of images with corresponding human attention to use in a Visual Question Answering (VQA) task. In VQA, a model is given an image and a question in the form of text as input, and the model must generate an answer in response. With this dataset, they compare where in an image do VQA models focus on (machine attention) when compared to where humans look at (human attention), i.e. where humans would consider to be the most salient parts of an image. We discuss this paper because it is another work that has used human attention in a computer vision and NLP-based task.

Zhao et al. (2019) uses co-attention in video captioning where different regions within frames as well as frames as a whole are attended to in the captioning process rather than gaze tracking information. Yu et al. (2017) attempts to use real gaze in training attention modules that can be used for different downstream tasks; however, the results of their approach are not as good as using ground truth real gaze directly. Saab et al. (2021) relies on gaze fixations to aid in the task of medical image classification. Different from the cited papers, we use a soft-Dynamic Time Warping (sDTW) loss, regularises how the predicted gaze saliency is allocated temporally in the video clip. Another way our work is different relative to the cited papers is that we are working with a smaller training dataset than those approaches. An earlier work from our group proposes using eye gaze in training to predict the salient regions in abdominal plane frames (Droste et al., 2019).

### 1.3. Contributions

This paper proposes an original multi-modal deep learning method to caption routine fetal US scan videos using sonographer spoken words



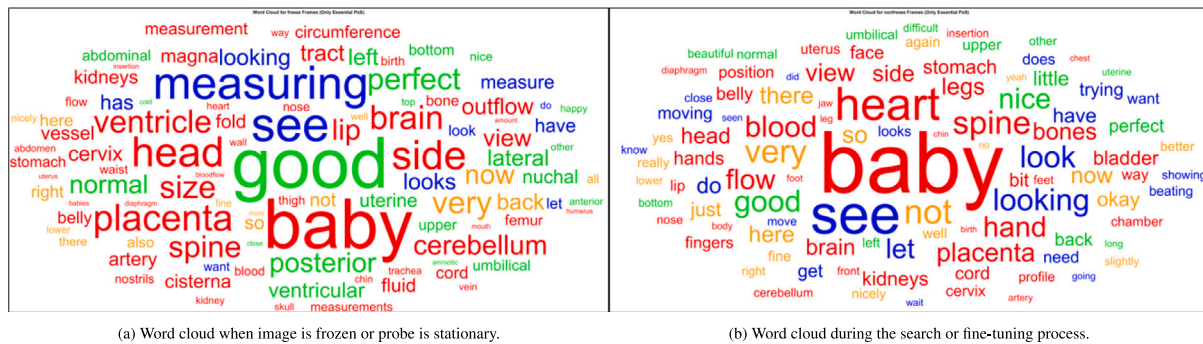


Fig. 1. Word clouds showing the diversity in the dataset. A larger word occurs more often. Red represents a noun, green represents an adjective, blue represents a verb, and orange represents an adverb. Words that make up the other parts-of-speech have been dropped for the sake of clarity and to emphasise more meaningful words that come in the form of nouns, adjectives, verbs, and adverbs.

derived from their speech recordings, and gaze-tracking data recorded during the scan acquisition process. This work is novel because: (1) it is the first attempt to perform automatic video captioning on fetal ultrasound scans using sonographer spoken vocabulary; (2) it proposes a novel mechanism to use expert gaze information. In terms of a quantitative evaluation metrics, the video caption generating model with gaze is found to perform better than one without. Furthermore, we are not aware of prior work that utilises gaze-tracking information to generate text for a medical imaging modality or, more generally, any work that attempts medical video captioning.

## 2. Methods

### 2.1. Multi-modal data preprocessing

We use a real-world dataset of second-trimester fetal ultrasound scan videos that were acquired as part of the PULSE project (Drukker et al., 2021). The PULSE project data includes mid-pregnancy ultrasound scans, undertaken between 18 to 22 weeks of gestation to screen for major fetal anomalies and growth abnormalities. These second trimester scans are offered in most high income countries and in the UK are carried out in accordance with the Fetal Anomaly Screening Programme (FASP) (NHS, 2018). The sonographers performing these scans follow a list of structures to be acquired and in the current analysis, we were interested in and chose to focus on these key fetal structures: head, abdomen, heart, and spine; therefore, the dataset can be categorised into four classes. The scan videos are recorded at 30 frames per second, and we centre the clips used in our analysis around freeze frames, which are detected using optical character recognition.

In the following subsections, we describe the multi-modal data preprocessing steps.

#### 2.1.1. Video, gaze, and audio capture

Routine fetal ultrasound videos and simultaneously acquired gaze data and speech recordings were captured for two sonographers giving a total of five routine full-length second-trimester (20 week) scans (Drukker et al., 2021). Videos and gaze data were acquired for another five scans that had retrospectively recorded audio, leading to a total of 10 full-length scan videos from the same number of women. One of the sonographers whose audio was collected during the scan was also the one from whom we obtained retrospectively recorded audio. A commercial Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machines<sup>3</sup> equipped with standard curvilinear (C2-9-D, C1-5-D), and 3D/4D (RAB6-D) probes were used to perform all the ultrasound scans used in our work. This study was approved by the

UK Research Ethics Committee (Reference 18/WS/0051), and written informed consent was given by all participating pregnant women. Each full-length scan recording has an average length of approximately 32 min (with standard deviation of approximately 14 min), which led to a total of approximately 320 min of audio–visual–gaze content. Gaze tracking data (x-y coordinates and time stamps) of sonographers were acquired at 90 Hz using an eye tracker (Tobii, Sweden).

The simultaneously acquired speech was recorded with a SHURE MX 184 Lapel Microphone. The microphone was connected to the PC in the scan room through the XLR-to-USB signal adapter. The acquired video and gaze information are also stored onto this device. All the collected data was then transferred from the device to our server. Audio is not recorded for the first 90 s after commencing a scan, allowing the sonographer and those present to exchange personal information not pertinent to our research. The raw audio is monaural in nature.

Video captions were obtained after transcribing sonographer speech recordings. The five retrospectively acquired audio recordings were transcribed through the Google Cloud Speech API (Google Cloud, 2019), while the rest containing significant conversation were transcribed manually in order to avoid the accidental transcription of non-sonographer speech consisting of the speech of the pregnant women or accompanying persons (Alsharid et al., 2019).

Clip–caption pairs were defined as follows. During a fetal US scan acquisition, the sonographer performs search, fine-tuning and interpretation of different fetal anatomies. The sonographer can freeze at a frame where they view the anatomy. Based on the freezing action, a video clip corresponding to every unique caption with the centre frame as a freeze frame was extracted. In Sharma et al. (2021), fetal ultrasound clips are also extracted from scan videos with respect to freeze frames. Then, we sample 12 frames from the segment of the clip spanning the length of the caption. For shorter clips, the 12 frames covered the entirety of the clip. For longer clips, the 12 frames are sampled from centremost segment of the clip in order to ensure a consistent amount of sampled video content. Fig. 2 shows a clip–caption pair.

The nature of the words in the dataset can be seen in the word clouds in Fig. 1. In image captioning tasks developed for spatio-temporal data (Alsharid et al., 2019), such as videos and speech recordings, each caption can correspond to more than one image, as spoken text naturally covers a temporal segment of visual content. In the current paper, we are not so constrained and have a unique caption for each video clip. The average length of a clip–caption pair in our dataset is 7.25 s. In general, in routine fetal US scans, sonographers spend unequal amounts of time looking at different anatomies of interest, which was also observed in our manually labelled video clip dataset (Sharma et al., 2019, 2021). This leads to a natural class imbalance for the four common anatomical classes considered in this work, namely, abdomen, head, heart, and spine. The distribution (% of clip–caption pairs) of the four anatomies is 11.7%, 31.4%, 40.5%, and 16.4% respectively.

<sup>3</sup> More information on this machine can be found here: <https://www.gehealthcare.co.uk/products/ultrasound/voluson-e8>.

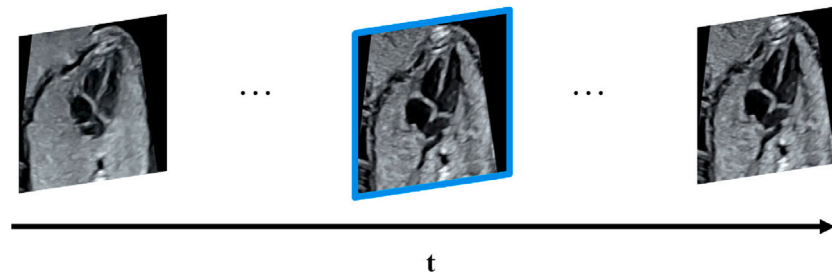


Fig. 2. Video clip of a beating fetal heart with the centred frame outlined in blue with the corresponding caption: “you can see the heart beating very nicely”. The original video clip is 6 s which means the clip includes around 180 frames. 12 frames from which will be sampled and used in training.

### 2.1.2. Gaze-tracking

The gaze data was filtered following the protocol described in Cai et al. (2018a). Binary maps were created from gaze data, with gaze-points labelled as 1 and others 0. Sonographer visual attention maps were subsequently generated by convolving the binary map with a Gaussian kernel with  $\sigma = 40$  pixels, assuming an observer-to-screen distance of 0.5 m, human field of view of  $1.5^\circ$  visual angle, and screen dimensions of  $33.2 \text{ cm} \times 20.7 \text{ cm}$ . The visual attention map was further normalised so that each pixel value is in the  $[0, 1]$  range. Predicted gaze saliency maps are obtained from a *spatio-temporal saliency prediction model* (Cai et al., 2020). In Cai et al. (2020), 10 frames are sampled from a video clip on which to predict the saliency of. For this work, this spatio-temporal saliency prediction model was trained using the same large simultaneous gaze-tracking dataset of second-trimester scans (Cai et al., 2020) with the aim, however, of predicting saliency maps of 12 sampled frames, rather than ten, from a video clip. Video clips from the PULSE data as well as the accompanying real gaze data was used in training the spatio-temporal saliency prediction to generate saliency maps. These generated saliency maps are what we refer to in this paper as predicted gaze. During training the spatio-temporal saliency prediction model, each video clip as well as the frames that constitute it had an anatomical label. Each frame also had a corresponding sonographer visual attention map.

All sonographers in the UK need to follow the FASP guidelines as mandated by the NHS; therefore, despite the existence of natural variation between them they are still required to scan and observe the same anatomical structures and perform the compulsory checks. Visual attention characteristics between individuals are different. However, in this scenario, given that the anatomy plane being searched for (abdomen, brain, heart or spine) is the same, the visual attention general falls within the same anatomical landmarks of interest (Drukker et al., 2020c). The nature of the work is to use gaze to assist in the generation of captions, and while the style of searching can be different between sonographers, the areas of interest on the fetus remain the same. Generally, the sonographers will look at the same regions on the fetal image, once the relevant anatomical plane is in view. Therefore, the generated captions are correct, since the same anatomical landmarks are present, the sonographer will be able to verbally affirm their appearance and existence.

For example, for the spine, the sonographers have demonstrated two different styles of scanning. One is where the sonographer glances at the spine briefly. The other is where the sonographer looks along the spine. In both cases, the sonographer’s gaze still falls on the spine even though their approach is different.

## 2.2. Modelling and analysis

### 2.2.1. Video captioning model architecture

We summarise the video captioning neural network architecture in Fig. 3. The architecture is inspired by the split-based captioning method (Tanti et al., 2017, 2018; Alsharid et al., 2019). The multi-modal deep learning framework has been designed for spatio-temporal

modelling of videos, further assisted by sonographer visual attention. Three variants of the model are considered: the gaze-less (GL) configuration, the real gaze (RG) configuration, and the predicted gaze (PG) configuration. The GL configuration, as its name suggests, has not been trained with gaze data, and the part of the framework that handles gaze information is not relevant to it. The RG configuration uses attention maps that come directly from the real gaze points of the performing sonographers as explained in more detail in this section. When training the model, the PG configuration uses saliency maps that have been generated by the previously trained saliency prediction model.

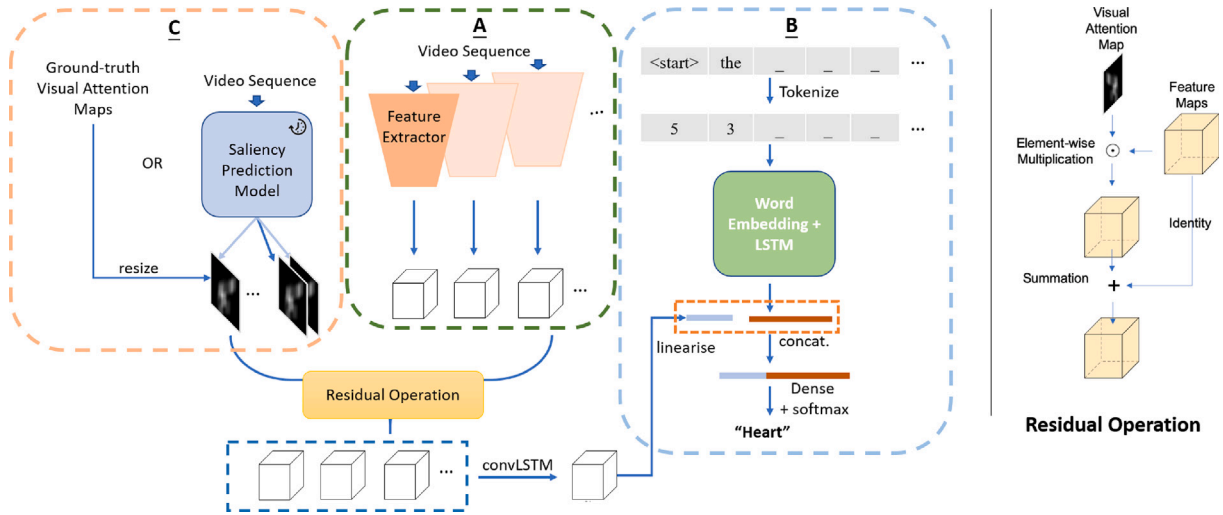
We explain the key features of Fig. 3.

**Block A** depicts the *visual spatio-temporal branch*, where each video clip is encoded by first extracting successive features of each constituent frame from the last convolutional block of a fine-tuned VGG-16 network (Simonyan and Zisserman, 2014), and then feeding these as a sequence into a convolutional-LSTM layer (Xingjian et al., 2015) to model the spatio-temporal context in the video clip. The last hidden state of the convolutional LSTM layer is flattened before going through a fully-connected layer.

**Block B** presents the *captioning branch* for learning joint video and text embeddings. Here, the tokenized sequence of a partial caption is passed through an embedding layer that represents each word with a 300-dimensional embedding vector. The sequence of embedding vectors is the input to an LSTM layer, the final hidden state of which is concatenated with the vector from the visual branch. Finally, a softmax operation is applied to the concatenated vector to output a probability distribution over the training vocabulary from which the next generated word is determined. The GL configuration is effectively the model architecture with Blocks A and B (but without Block C). The RG model configuration and the PG model configuration all include Blocks A, B, C. They differ in that in the PG model, in Block C, the maps come from the saliency prediction model. Whereas, in the RG model configuration, the maps are the ground truth visual attention maps obtained through real sonographer eye gaze points.

The selection of this architecture, illustrated in Fig. 3 as **Block A** and **Block B**, is primarily motivated by the fact that we have a relatively small sized dataset to work with. The captioning branch in Block B shows that text encoding LSTM will not encounter the image information, making it easier to train the LSTM-RNN as it would only need to learn the text information. There is benefit to doing this for image captioning when working with a relatively small sized dataset (Tanti et al., 2017, 2018). **Block A** transfer learns from the VGG-16 network pretrained on ImageNet (Deng et al., 2009) and fine-tuned on fetal ultrasound images. We used that same CNN only duplicated twelve times, once for every sampled frame, in the sequence before being fed into a convolutional LSTM to process and handle changes in temporal information between those twelve frames. More information on fine-tuning the VGG16 can be found in Appendix A.2 in Appendix.

The *gaze-assisted model* configurations that include RG or PG have the gaze information included through a *gaze-encoding branch* shown in **Block C**. In both cases, the visual attention maps from Block C



**Fig. 3.** The architecture of the multi-modal model is shown in this figure. The specifics of the residual operation that is performed on the outputs of Blocks A and C are shown on the right side of this figure. The gazeless model configuration only includes Blocks A and B. Block A represents the branch of the model where the spatial feature information is extracted from the video clip for each of its sampled frames by a VGG-16 network. Block B represents the branch handling text information. The sequence of words generated so far are tokenized and embedded with a Word2vec embedding vector before being passed as input to an LSTM-RNN. The last hidden state of this LSTM-RNN is concatenated with the linearised feature vector from the convolutional LSTM. The real gaze and predicted gaze model configurations also include Block C. In Block C, either the ground truth visual attention maps are used (in the case of the real gaze configuration) or predicted gaze saliency maps are predicted and then used for each sampled frame in a video clip by a previously trained saliency prediction model. The extracted feature blocks from Block A and the gaze maps from Block C are combined together through the residual operation shown in detail on the right side of the figure. In the gazeless configuration, there is no residual operation to be performed on the sequence of feature blocks. They are passed directly to the convolutional LSTM. A more detailed version of this architecture is shown in Fig. 9 in Appendix.

filter the features extracted from Block A using a residual operation, i.e. element-wise multiplication followed by identity summation to avoid the vanishing gradient problem *in situ* while highlighting visually-salient regions in image features. Eq. (1) shows the residual operation

$$\mathbf{R} = \mathbf{V} \odot \mathbf{F} + \mathbf{F}, \quad (1)$$

where  $\mathbf{R}$  is the result of the residual operation,  $\mathbf{V}$  is the visual attention map duplicated enough times to make it possible to perform element-wise multiplication with the features, and  $\mathbf{F}$  represent the feature maps extracted from a CNN.

In the RG configuration, ground-truth real gaze visual attention maps are used as the visual attention maps. They are directly resized to match the dimension of the features extracted from each frame in Block A.

In the PG configuration, input US images in Block A are fed through a spatio-temporal variant of a pre-trained saliency prediction model (Cai et al., 2020), which we describe below, to predict visual attention maps. This saliency prediction model is spatio-temporal because it accounts for changes in the sonographer's attention throughout a video clip.

### 2.2.2. Saliency prediction model

The spatio-temporal saliency prediction model first models “static visual attention” using the method of Cai et al. (2018a) and predicts a visual attention map on each input video frame. Then, the features extracted are fed into a bi-directional convolutional-LSTM to model “dynamic visual attention”, accounting for the temporal variation of sonographer visual attention. Using the bi-directional convolutional LSTM makes it possible to obtain meaning from the spatio-temporal information in these fetal ultrasound videos that make up the dataset.

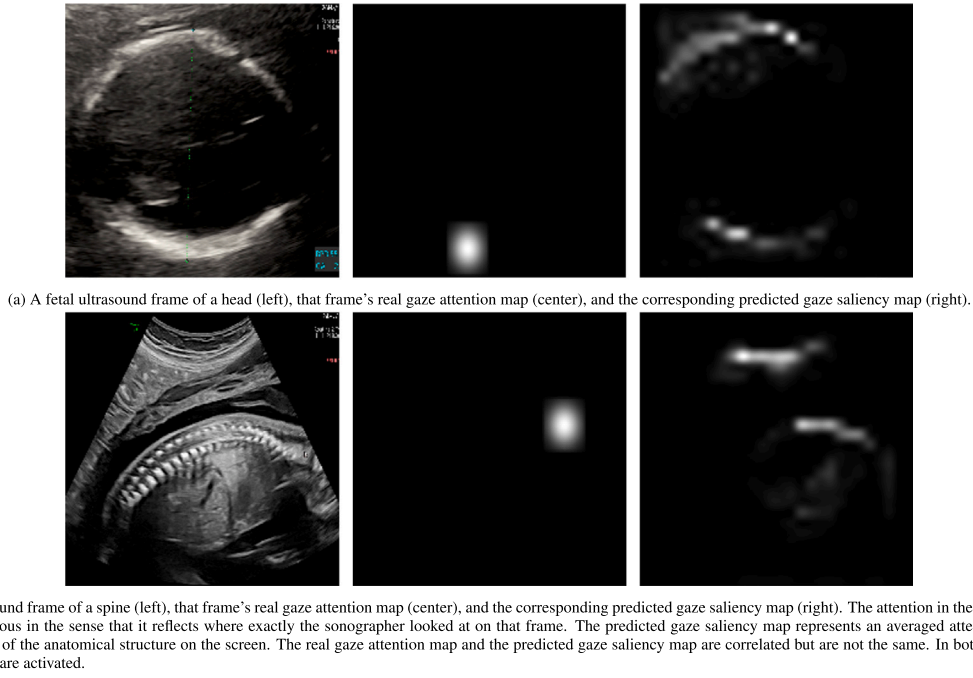
The spatio-temporal saliency prediction model is trained through three losses. In addition to the *mean squared error (MSE)* and *Kullback-Leibler Divergence (KLD)* losses, the model employs a soft-Dynamic Time Warping (sDTW) loss (Cuturi and Blondel, 2017) to align predicted visual attention maps and ground-truth maps by regularising how the predicted visual attention is allocated temporally in the target video clip. The saliency prediction model consists of two modules, a temporal

attention module and a video classification module. The video classification module is used when training the saliency prediction model before using it as part of the gaze captioning framework. CNN-extracted feature maps are fed into the temporal attention module in both forward and reverse order. Extracted spatial features are passed through several convolutional layers to generate a static attention map. The static attention map also undergoes the residual operation, allowing us to obtain a modified map. The modified map from both the forward and reverse orders are each inserted into a convolutional recurrent neural network. The hidden states from both are concatenated together. The concatenated blocks undergo an additional convolution operation followed by sigmoid activation, thereby obtaining the predicted gaze saliency maps which are then used in the PG configuration of the captioning model. Fig. 4 shows a couple of representative fetal ultrasound frames and their corresponding real gaze attention maps and predicted gaze saliency maps.

### 2.2.3. Model training

Cross-entropy loss is among the losses commonly used in NLP tasks (Goldberg, 2016) and captioning specifically (Tanti et al., 2017, 2018; Alsharid et al., 2019). Focal loss is a modified version of cross-entropy loss that gives a greater importance to misclassifications by down-weighting correct classifications when training a deep learning model, hence, it is useful for classification problems that have imbalanced input datasets (Lin et al., 2017). In our training vocabulary, word imbalance is observed. For example, ‘the’, being an essential article used in the English language, is unsurprisingly one of the most commonly occurring words in our dataset, having over 200 instances. On the other hand, ‘iliac’ is a word that is significantly more relevant anatomically but exists in only one caption. However, the more common words, despite their grammatical importance, are not as essential for the anatomical description as some of the less-represented words. In practice, without addressing natural imbalance, a model is more likely to generate the more common words due to their prevalence. As captioning can be considered as a classification problem with words as classes, the vocabulary imbalance justifies choosing focal loss in this work.





**Fig. 4.** Two fetal ultrasound frames, their real gaze attention maps, and their corresponding predicted gaze saliency maps using the method of Cai et al. (2018). This figure shows the information the different model configurations (gazeless, real gaze, predicted gaze) learn from.

Eq. (2) defines the focal loss function  $FL$

$$FL = \sum_{w=1}^v -\alpha (1 - p_w)^\gamma y_w \log(p_w) \quad (2)$$

where  $\alpha=0.25$  and  $\gamma=2$  and where  $w$  is an index in a word list,  $v$  is the total number of words in the vocabulary,  $y_w$  is the associated ground truth value of that word, and  $p_w$  is the associated softmax probability of that word.  $\alpha$  is described as a weighting factor and  $\gamma$  is the focusing parameter. We use the same values for  $\alpha$  and  $\gamma$  ( $\alpha=0.25$  and  $\gamma=2$ ) as in Lin et al. (2017).

To augment the original training dataset, the sequence of frames was randomly augmented on the fly during training by either rotating by an angle between  $-30^\circ$  and  $30^\circ$  around the vertical  $y$ -axis or by horizontally reflecting the image. We used Adam optimisation (Kingma, 2015) with a learning rate of 0.01. Early stopping was done when the validation loss ceased to improve after 28 epochs. To reduce overfitting, dropout with a rate between 0.4 and 0.5 was used on the output of the LSTM and after merging the text and image information. The different model configurations (see Section 2.2.1) were trained with teacher forcing (Goodfellow et al., 2016). Teacher forcing is a commonly used practice to train recurrent neural networks with sequential information. It involves providing a sub-sequence of ground truth words of appropriate length as input to the network at every training time step, regardless of what output (sequence of words) has been generated by the model in the previous time steps (Goodfellow et al., 2016). In our case, if a ground truth caption is ‘the healthy spine’, and we are at timestep  $t = 2$ , the partial caption that will be provided as input to the model at that timestep is [‘start-token’, ‘the’, ‘healthy’] which come from the ground truth caption. The model being trained, though, may have generated different predicted words in the previous timesteps. For example, it may have generated ‘a’ at  $t = 0$ , and ‘curved’ at  $t = 1$ , but [‘start-token’, ‘a’, ‘curved’] will not be used as input to the model at  $t = 2$ . This training behaviour continues at all other timesteps. At inference, the model relies on its previously generated words and the corresponding sequence of frames to generate the next word. We use a framework where there is a captioning model for each of the four anatomical structures. A video clip classifier based on Block A classifies the video clip and then starts the appropriate captioning model based on the classification output.

**Table 2**

Results of model configurations with different word embeddings.

Config.	B1	B2	B3	B4	RL	Rich.	F1	ARS
Word2vec	<b>0.26</b>	<b>0.19</b>	<b>0.16</b>	<b>0.09</b>	0.35	<b>0.11</b>	<b>0.78</b>	<b>0.36</b>
BioWV	0.13	0.09	0.08	0.04	<b>0.44</b>	0.09	0.75	0.30

#### 2.2.4. Word embeddings

Word2vec is an established embedding method by which words are represented by embedding vectors (Mikolov et al., 2013a). It is effectively a neural network with a single hidden layer trained to predict either the next word given previous words or the context, as the target, around the word given as input (Mikolov et al., 2013b,a). BioWordVec is a vector representation for words that are specifically relevant to the biomedical domain (Zhang et al., 2019). It is based on the fastText embedding (Bojanowski et al., 2016) but goes beyond by incorporating Medical Subject Headings (MeSH) terms, which is effectively in the form of an ontology (Zhang et al., 2019). On the other hand, Word2vec is not specifically optimised to make use of medical text and information. The key improvement of fastText (Bojanowski et al., 2016) over Word2vec is that fastText takes the subunits of words into consideration whereas Word2vec does not. One could see how that could be useful in the biomedical domain, where even the word ‘biomedical’ itself could be divided into the sub-units, ‘bio’ and ‘medical’; however, truly noticing this advantage depends on having a dataset consisting of enough words that can be divided into these sub-units (Zhang et al., 2019). In this work, we compare the Word2vec and BioWordVec embeddings within our modelling framework in an experiment reported in Section 3.1. We used pre-trained embedding vectors, but the embeddings were fine-tunable during the training of the captioning model.

#### 2.3. Evaluation metrics

We use two sets of metrics to evaluate modelling performance; one focused on the quality of video clip classification and one focused on the quality of the generated caption. With regards to the visual content classification-related metrics, we calculate, at the clip level, the

**Table 3**

Results of model configurations with varying levels of gaze involvement. There are three rows, one for each the model configurations that we experiment with: the gazeless (GL) model, the model that uses real gaze (RG) attention maps, and the model that used predicted gaze (PG) saliency maps. The columns show the different evaluation metrics with which we compare the different model configurations. B1-B4, RL, and Rch. are more caption-focused evaluation metrics, while F1 and ARS are class-focused. More information on the different evaluation metrics can be found in Section 2.3 and the Appendix. The scores in bold in each column are the ones that are highest for that evaluation metric.

	B1	B2	B3	B4	RL	F1	ARS	Eff	Rch
UL	<b>0.33</b> (0.01)	<b>0.17</b> (0.00)	<b>0.12</b> (0.00)	<b>0.08</b> (0.00)	<b>0.50</b> (0.00)	0.22 (0.01)	–	0.05 (0.00)	0.02 (0.00)
GL	0.27 (0.02)	0.12 (0.01)	0.07 (0.01)	0.03 (0.00)	0.43 (0.02)	0.77 (0.12)	<b>0.24</b> (0.05)	0.17 (0.01)	0.05 (0.00)
RG	0.28 (0.01)	0.10 (0.01)	0.05 (0.01)	0.02 (0.00)	0.44 (0.01)	0.81 (0.08)	0.18 (0.03)	0.19 (0.01)	0.05 (0.00)
PG	<b>0.33</b> (0.01)	0.15 (0.01)	0.08 (0.01)	0.04 (0.00)	0.48 (0.01)	<b>0.91</b> (0.03)	0.23 (0.03)	<b>0.25</b> (0.01)	<b>0.08</b> (0.00)

**Table 4**

Results of model configurations with varying levels of gaze involvement being shown for each anatomical structure specifically. The best score for the combination of a particular structure and a particular metric is in bold.

	Structure	B1	B2	B3	B4	RL	F1	ARS	Eff	Rch
GL	Abdomen	0.27	0.08	<b>0.03</b>	<b>0.00</b>	0.41	0.68	<b>0.26</b>	0.12	0.05
	Head	0.27	0.15	0.11	0.06	0.40	0.77	0.30	0.17	0.06
	Heart	0.34	0.15	<b>0.09</b>	<b>0.06</b>	0.45	0.87	<b>0.21</b>	0.21	0.03
	Spine	0.21	0.09	0.04	0.02	0.48	0.76	0.19	0.20	0.07
RG	Abdomen	0.30	0.07	0.02	<b>0.00</b>	0.42	0.75	0.20	0.14	<b>0.06</b>
	Head	0.30	0.16	0.13	0.07	<b>0.46</b>	0.98	0.27	0.22	0.07
	Heart	0.33	0.12	0.05	0.02	0.46	0.80	0.12	0.23	0.03
	Spine	0.20	0.06	0.01	0.00	0.42	0.70	0.11	0.17	0.06
PG	Abdomen	0.35	0.11	0.00	0.00	0.47	<b>0.79</b>	0.15	<b>0.15</b>	<b>0.06</b>
	Head	<b>0.35</b>	<b>0.21</b>	<b>0.17</b>	<b>0.10</b>	0.44	<b>1.00</b>	<b>0.40</b>	<b>0.32</b>	<b>0.14</b>
	Heart	0.36	<b>0.16</b>	0.09	0.04	<b>0.51</b>	<b>0.92</b>	<b>0.21</b>	<b>0.29</b>	<b>0.05</b>
	Spine	0.27	0.12	0.05	0.03	0.51	<b>0.92</b>	<b>0.17</b>	<b>0.22</b>	<b>0.08</b>
UL	Abdomen	<b>0.44</b>	<b>0.16</b>	0.00	<b>0.00</b>	<b>0.52</b>	0.33	–	0.05	0.02
	Head	0.30	0.16	0.13	0.07	<b>0.46</b>	0.36	–	0.07	0.02
	Heart	<b>0.39</b>	<b>0.16</b>	<b>0.13</b>	0.07	<b>0.49</b>	0.00	–	0.00	0.00
	Spine	<b>0.35</b>	<b>0.21</b>	<b>0.18</b>	<b>0.12</b>	<b>0.54</b>	0.33	–	0.08	0.04

classification F1-Score and a specifically designed anatomical relevance score (ARS) that was introduced in Alsharid et al. (2019). Caption quality is reported using BLEU (Papineni et al., 2002), and ROUGE-L (Lin, 2004). We also report metrics that measure caption efficiency and richness. Let  $\Gamma$  represent the generated words, and  $Y_\alpha$  represent the words that are relevant with respect to an anatomical structure  $\alpha$ . The intersection of  $\Gamma$  and  $Y_\alpha$  gives the generated relevant words. The Efficiency  $\epsilon$  is defined by

$$\rho = \frac{\#(\Gamma \cap Y_\alpha)}{\#(Y_\alpha)}. \quad (3)$$

It has a high value for correct captions with a few words. The Richness  $\rho$  is defined as

$$\epsilon = \frac{\#(\Gamma \cap Y_\alpha)}{\#(\Gamma)}, \quad (4)$$

and intuitively reflects the completeness of a generated caption.  $\#()$  represents a count function that returns the number of elements contained within.

This relevance has to do with what words have been deemed relevant to an anatomical structure. For example, the ‘head’ anatomical structure will consist of words that are of relevance to it. These words include ‘nuchal’, ‘cerebellum’, and ‘skull’. The number of relevant words is included in the calculations of both richness and efficiency, but the way they are used is different.

Richness and Efficiency are inspired by precision and recall. However different from precision and recall, a value of 20%–30% would be good for efficiency (ground truth captions get efficiency scores in that range). In a caption consisting of five words, this would mean at least one is of appropriate anatomical meaning. If we were using traditional precision and recall metrics, all five words would have to be relevant in order to get a high score, but that does not make sense lexically. Therefore, the scores are naturally low, because the words that we consider to be anatomically relevant are, by design, a small number relative to the size of the entire vocabulary.

Traditional metrics, such as BLEU and ROUGE reward models for generating appropriate captions that align well with the ground truth irrespective of the actual semantic content. For example, a model might learn to generate sentences in the same style as the sonographers (the use of determiners and verbs are similar); therefore, there will be overlap with the ground truth. For instance, there will be high BLEU and ROUGE score between these two sentences: ‘we can see the heart beating’ and ‘we can see the curved spine’ because there is significant overlap between the two sentences, but we can understand that semantically the two sentences are talking about two different structure entirely. This is what makes alternative metrics, such as ARS, richness, and efficiency useful. Ideally, you want both the traditional metrics and the semantics focused metrics to be high, but it is more important for the content to be relevant than it is for the sentence to look more similar to the ground truth. This is a strong argument for relying on the semantics focused metrics. There are many possible ways to form a sentence that is semantically meaningful but differs in structure from the ground truth.

### 3. Results

#### 3.1. Quantitative evaluation

Table 2 compares results when using Word2vec, and when using BioWordVec (BioWV) with the RG configuration. The model configuration trained with word2vec mostly outscored the model configuration trained with BioWordVec.

We compared the different model configurations that involve eye tracking data using leave-one-out cross-validation. There are nine different runs, where video clips from eight different scans go into the training set and video clips from a ninth scan make up the validation set. The tenth video that is not part of the cross-validation experiment is set aside to later be used to independently evaluate the trained models.

Table 3 shows the scores obtained with the three multi-modal configurations. GL, RG, PG are the gaze-less, real gaze, and the predicted gaze configurations respectively. B1 to B4, RL, F1, ARS, Eff, and Rich. represent BLEU-1 to BLEU-4, ROUGE-L, the F1 score, the Anatomical Relevance Score, Efficiency, and Richness, respectively, with standard deviations in brackets.

The results shown in Table 3 are those after leave-one-out cross validation has been performed. For every single fold and model configuration, there are four models that are trained (one for each anatomical structure). As Table 3 shows, caption generation assisted by sonographer gaze improves results, giving the overall best performance”.

Our gaze-assisted configurations outperform our gaze-less configuration on richness, implying a more thorough use of relevant terminology when incorporating gaze information in the captioning models. The PG configuration outperformed the RG configuration. Table 4 breaks down the results of Table 3 by showing the scores for each anatomical structure.

We have also included a comparison with the Transformer-based UniVL (Luo et al., 2020) (referred to as UL in Tables 3 and 4). We compare with UniVL because it is one of the current best-performing models for captioning and other related tasks with publicly available code,



enabling a direct comparison. It is the current SOTA model<sup>4</sup> for the YouCook2 dataset (Zhou et al., 2020). UniVL had been pretrained on the HowTo100M dataset (Miech et al., 2019). We fine-tuned it on ours following the recommended approach of the authors. The pretraining gives UniVL an enormous advantage in the formulation of syntactically sound sentences; however, the generated captions obtained low scores on *F1*, *Efficiency*, and *Richness*. All three incorporate a sense of whether the generated caption includes words anatomically relevant to those in the ground truth one. We attribute this result to the nature of our dataset which is considerably smaller compared to the datasets (shown in Table 1) that would be used for benchmarking purposes by UniVL and similar models. Figs. 7 and 8 show this observation in a qualitative fashion.

#### 4. Discussion

In Table 2, we compare word embeddings. It is interesting to note that BioWordVec underperforms compared to Word2vec in all but one metric. This result is explainable by the fact BioWordVec is focused on biomedical terms which does not constitute the majority of our data. 43.3% of the spoken words in our data are determiners, prepositions, pronouns, adverbs, and other related parts-of-speech (Drukker et al., 2021). Transcribed sonographer speech is different in structure and content from PubMed articles which BioWordVec is shown to be suitable for by Zhang et al. (2019). Zhang et al. (2019) also determined that using MeSH terms with BioWordVec did not significantly help in NLP tasks involving clinical notes, but they might be effective with NLP tasks involving text from PubMed articles. Despite the medical context in which the data is acquired, the nature of our data is a conversational between an expert and a layperson. Conversational words like ‘baby’ and ‘belly’ are more common than ‘fetus’ and ‘abdomen’ with biomedicine-related words such as the names of proteins or medications being entirely absent. Rudimentary medical information is communicated within a medical context to a layperson in a way that they could understand in conversational language. The experimental results show that Word2vec is more suitable for transcribed speech than BioWordVec.

The modelling problem in this work is different from prior work (Alsharid et al., 2019, 2020, 2021). Alsharid et al. (2019, 2020, 2021) solve the simpler problem of image captioning (one image plus associated text, no gaze). Because they deal with a different modelling task and the actual data used in those papers, although still under the same project, is slightly different. Therefore, a direct comparison cannot be made. We also do not compare with the curriculum learning method of Alsharid et al. (2020) and Alsharid et al. (2021) because the contributions of those works cannot be directly compared with the current one, since we do not introduce a data preprocessing and batch preparation strategy in the current paper (see Figs. 5 and 6). It is interesting to observe the high scores for the predicted gaze configuration obtained when compared to the real gaze configuration in Table 3. We reason that this is so since with predicted gaze, all frames would have attention maps that they could leverage from. With real gaze, on the other hand, only frames with real sonographer gaze points will have an attention map generated for them. Some frames do not have any gaze points because the sonographer may have been looking at the subject or they were staring at a part of the screen that was not showing the anatomical content or were not looking at the screen at all.

The BLEU scores are lower than reported for in other natural language processing tasks including image captioning (Bernardi et al., 2016; Vinyals et al., 2015). This result may be because only one ‘ground truth’ caption to compare generated captions with. Other natural image captioning datasets, such as MS-COCO (Lin et al., 2014), Pascal VOC

2008 (Farhadi et al., 2010), Flickr8k (Rashtchian et al., 2010), and Flickr30k (Young et al., 2014), have up to five different ‘ground truth’ captions for the same image (Vinyals et al., 2015). BLEU scores directly compare the degree of overlap between a generated caption and the reference. When compared to those of other NLP tasks, BLEU scores in this task are indeed relatively not high but are not enough to dismiss the generated captions as erroneous since the BLEU scores only have a single reference to compare them with. However, the incorporation of gaze improves the scores and qualitative performance. The PG-trained model scores highest among our configurations. We attribute this result to the fact that with PG-trained models, the attention is focused on objects (or parts of objects), so the saliency in a head image would be depicted as around a head for example. This is in contrast to the input of the RG-trained models where the attention is localised to pixel regions based on gaze points. Another possible reason, as mentioned earlier, is that not all frames have real gaze points (from which to create real gaze attention maps), but all frames have predicted gaze saliency maps.

We intend to overcome limitations of this work in future extensions. These include using transformers in the fetal video clip captioning model architecture. We also intend to explore more sophisticated transformer-based word embedding models, such as BERT (Devlin et al., 2018), BioBERT (Lee et al., 2020), and BioALBERT (Naseem et al., 2021) to see how context aware embeddings models can potentially improve the performance. Extended vocabularies that better capture the essence of performed sonography actions and the adoption of more sophisticated spatio-temporal visual and textual feature extractors will be considered in future work.

#### 5. Conclusion

We have proposed an automatic video captioning method to describe spatio-temporal video content using words from the spoken vocabulary of professional sonographers in routine second-trimester fetal ultrasound scans. Word2vec embedding outperformed BioWordVec in gaze-assisted captioning. Utilising gaze in captioning helps achieve higher scores on evaluation metrics, specifically *BLEU-1* to *BLEU-4* and *F1*. Predicted gaze has the added benefit of allowing all frames to have accompanying gaze information that could help in the captioning process, while real gaze, although it better reflects the important anatomical content (being from the sonographer directly), has the downside of only being available for those frames where the sonographers happened to be looking at the screen. The proposed method can potentially be explored for other modalities with a temporal component.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Data availability

The authors do not have permission to share data.

#### Acknowledgements

We acknowledge the ERC (ERC-ADG-2015 694581 project PULSE), the EPSRC, United Kingdom (EP/MO13774/1), the Rhodes Trust, and the NIHR BRC funding scheme. We would like to thank Clare Teng and Mourad Gridach for participating in useful discussions.

<sup>4</sup> [paperswithcode.com/sota/video-captioning-on-youcook2](https://paperswithcode.com/sota/video-captioning-on-youcook2)

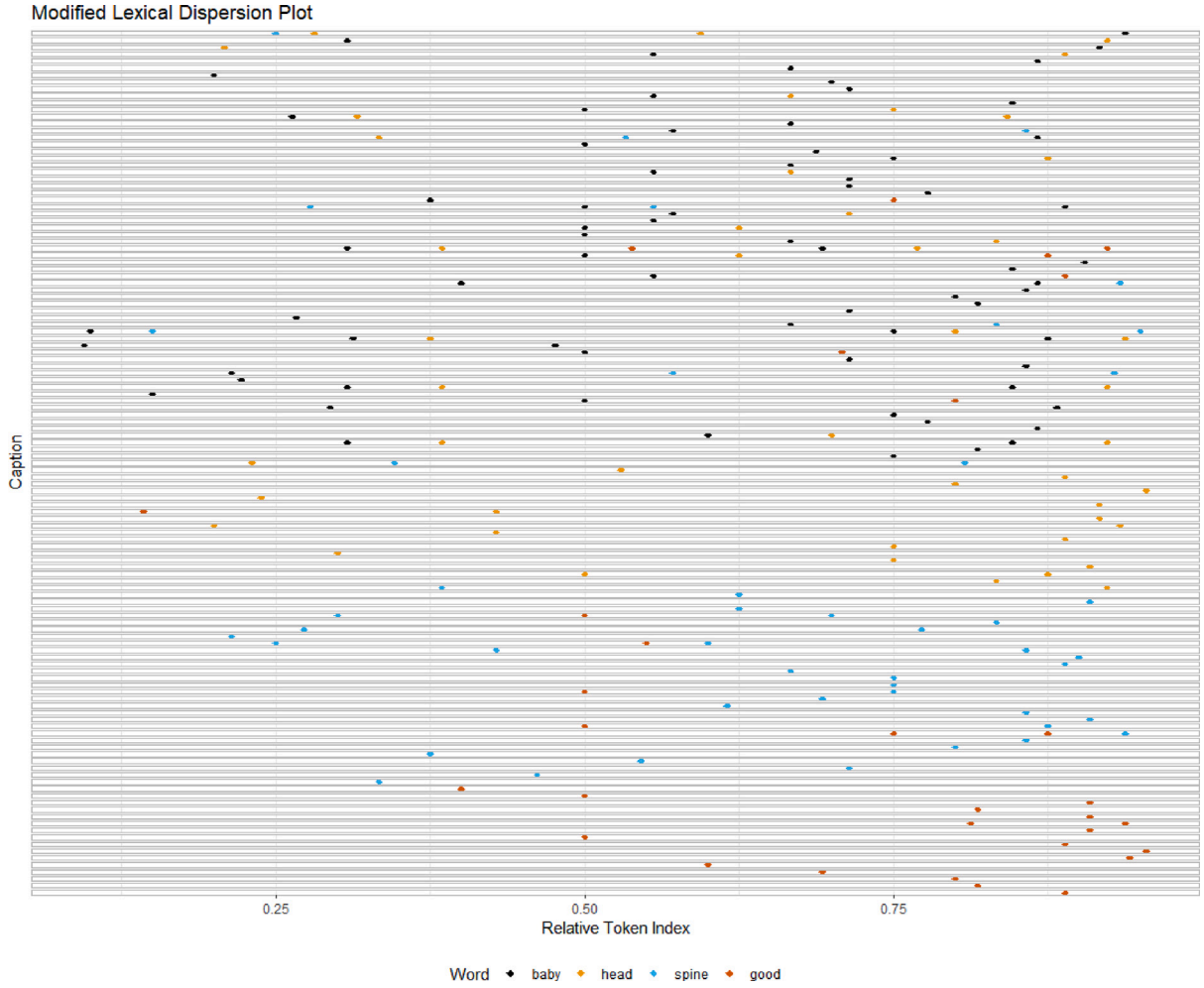


Fig. 5. A modified lexical dispersion plot showing where do four of the most common words exist in the captions of our dataset in relation to one another. Each row represents a caption. The caption lengths have been normalised for the purposes of this visualisation.

## Appendix

### A.1. Evaluation metrics

The F1-Score as described in this work uses Scikit-learn's (Pedregosa et al., 2011) implementation of the F1-Score where it is the weighted average of recall and precision. The ground target values represent the actual classes (anatomical structures) of the ground truth captions. The predicted target values represent the predicted classes (anatomical structure). To determine the anatomical structure of a caption, we had prepared a function that serves as a structure determiner. Given a caption, it returns the most likely anatomical class for that caption. The structure determiner works by finding which of the vocabularies of the anatomical structures has the most word overlap with the caption of interest. These vocabularies have been handcrafted and prepared with assistance from clinical experts. F1-Score is defined as:

$$F1 = 2 * \frac{\text{recall} * \text{precision}}{\text{recall} + \text{precision}}. \quad (5)$$

The highest possible value is one, and the lowest possible value is zero. The Anatomical Relevance Score (ARS) is also reliant on the structure determiner. Eqs. 6, (7), and (8) show how ARS is calculated (Alsharid et al., 2019) as follows.

$$CS_k = \left( \sum_{i=1}^{L(W^c)} \mathbf{1}_{V_k}(w_i^c) \right)^{-1} \sum_{i=1}^{L(W^c)} \mathbf{1}_{V_k}(w_i^c) p_i \quad (6)$$

$$SS_c = \begin{cases} \max_{k \in K} CS_k & \text{if } \arg\max_{k \in K} (CS_k) = GT_c \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$ARS = \frac{1}{C} \sum_{i=1}^C SS_c \quad (8)$$

Let  $CS_k$  be a quantitative scalar score that a caption has in relation to the anatomical structure (class)  $k$ ,  $K$  be the set of four anatomical structures (classes),  $V_k$  be the vocabulary (set of words) of class  $k$ ,  $L$  be the length of a caption  $W^c$  that consists of words  $w_i^c$  with softmax probabilities  $p_i$ ,  $\mathbf{1}_V(\cdot)$  be an indicator function which returns 1 if  $w_i$  is in  $V$  and 0 otherwise,  $SS_c$  be a quantitative scalar score that only considers  $CS_k$  if it has the ground truth anatomical structure (class)  $GT_c$ , and  $C$  be the total number of captions in the test set.  $C$  would be equivalent to the total number of image-caption pairs in the test set (Alsharid et al., 2019).

For each of the four anatomical classes, a  $CS_k$  is obtained. We have  $CS_{abdomen}$ ,  $CS_{head}$ ,  $CS_{heart}$ , and  $CS_{spine}$ . When calculating  $CS_k$  for an anatomical structure, if a word in the generated caption exists in the vocabulary of this anatomical structure,  $V_k$  will return one for that word, and therefore its softmax probability will be added, contributed to the score obtained for  $CS_k$ . So, all the words in the generated caption that are relevant to the vocabulary of that anatomical structure will be counted. These softmax probabilities are summed and then divided by the total number of relevant words, so that, ARS is not biased towards longer captions that happened to repeatedly generate the same relevant word multiple times.

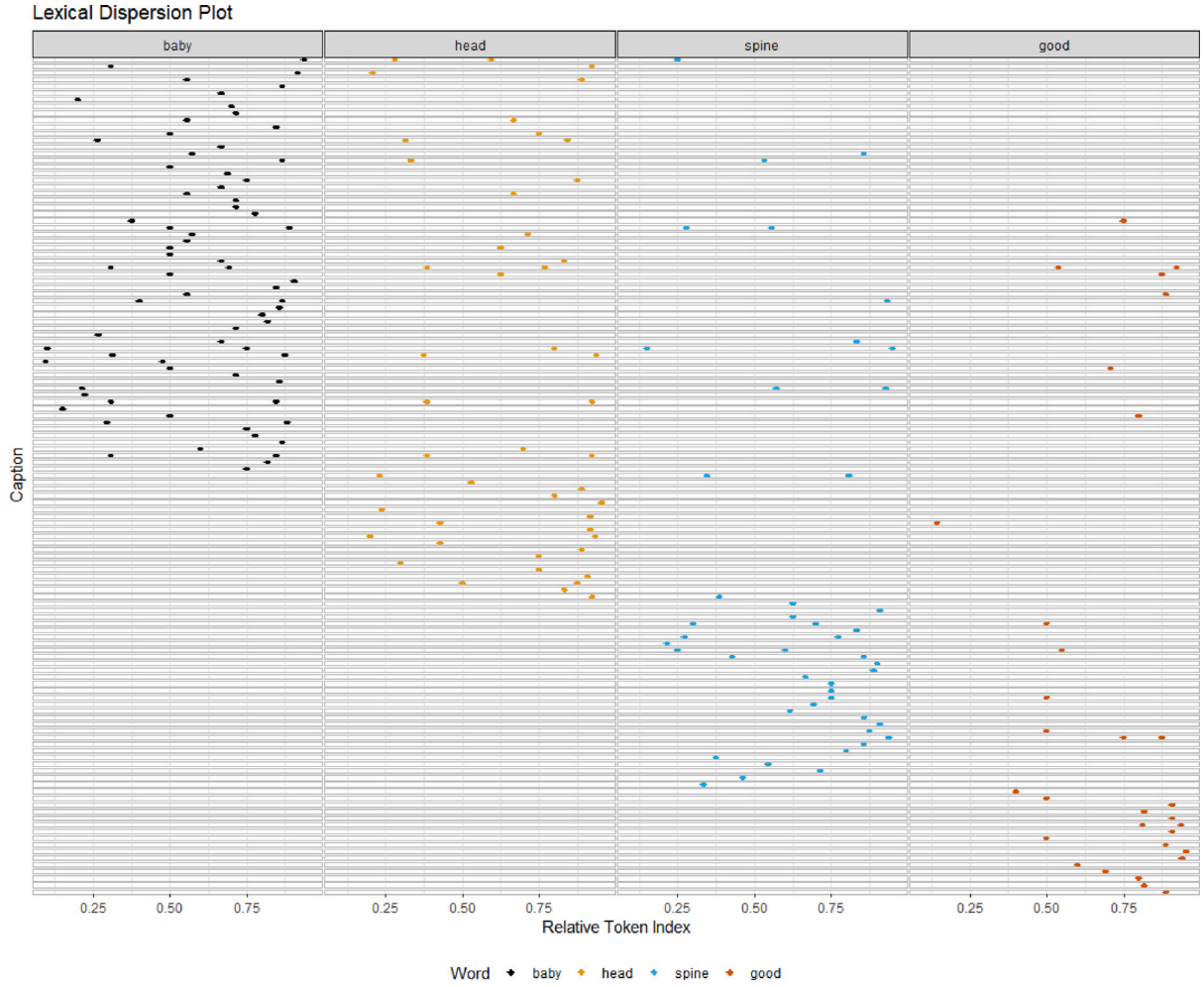


Fig. 6. A lexical dispersion plot showing where do four of the most common words exist in the captions of our dataset but with each word having its own facet. Each row represents a caption. The caption lengths have been normalised for the purposes of this visualisation.

After obtaining  $CS_k$  for each of the four structures, we determine which is the maximum in value. Afterwards, we can obtain  $SS_c$ . If the  $CS_k$  with the highest value was of the anatomical class that happened to be the ground truth anatomical class (i.e. the model generated a caption relevant to the right anatomical structure), then it will be assigned as the value of  $SS_c$ . Otherwise,  $SS_c$  will be given a value of zero.  $SS_c$  is the score associated with a single data sample from the test set.

Finally, after we have obtained  $SS_c$  scores for all the data samples in the test set, we sum them up and then divide by the total number of captions in the test, to obtain the  $ARS$  score.

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L) (Lin, 2004) look for overlap between generated captions and ground truth captions. BLEU will look for overlap at the level of one word or 1-gram (BLEU-1), two words or 2-gram (BLEU-2), three words or 3-gram (BLEU-3), and four words or 4-gram (BLEU-4). Eq. (9) shows how to calculate the BLEU scores. Eq. (10) shows how to calculate the ROUGE-L score between a generated (GEN) caption and the ground truth (GT). In Eq. (10),  $P$  is the precision,  $R$  is the recall,  $lcs$  is the longest common subsequence, and  $\beta$  is a constant parameter.

$$BLEU = brevity\_penalty * e^{\sum_{n=1}^N \frac{1}{N} \log \frac{\text{number of matched } n\text{-grams}}{\text{total number of } n\text{-grams}}} \quad (9)$$

$$ROUGE_L = \frac{(1 + \beta^2) R_{lcs}(GEN, GT) P_{lcs}(GEN, GT)}{R_{lcs}(GEN, GT) + \beta^2 P_{lcs}(GEN, GT)} \quad (10)$$

Table 5  
Quantitative results for Fig. 7(a).

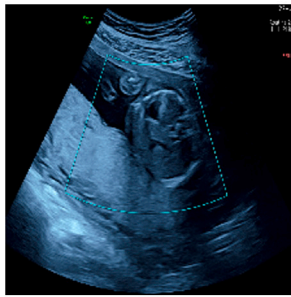
	B1	B2	B3	B4	R	F1	Eff	Rch
GL	0.15	0.09	0.00	0.00	0.32	1	0.21	0.25
RG	0.12	0.00	0.00	0.00	0.27	1	0.13	0.17
PG	0.12	0.00	0.00	0.00	0.27	1	0.13	0.17
UL	0.33	0.00	0.00	0.00	0.51	0.00	0.00	0.00

Table 6  
Quantitative results for Fig. 7(b).

	B1	B2	B3	B4	R	F1	Eff	Rch
GL	0.18	0.00	0.00	0.00	0.29	1.00	0.30	0.25
RG	0.22	0.09	0.00	0	0.38	1	0.13	0.17
PG	0.24	0.14	0.12	0.00	0.65	1	0.13	0.17
UL	0.43	0.30	0.24	0.00	0.68	0.00	0.00	0.00

Table 7  
Quantitative results for Fig. 7(c).

	B1	B2	B3	B4	R	F1	Eff	Rch
GL	0.39	0.25	0.18	0	0.49	1	0.17	0.09
RG	0.29	0.13	0.10	0	0.42	1	0.42	0.15
PG	0.22	0.14	0.10	0	0.42	1	0.39	0.15
UL	0.02	0.01	0.01	0	0.36	0	0	0



**GT:** belly

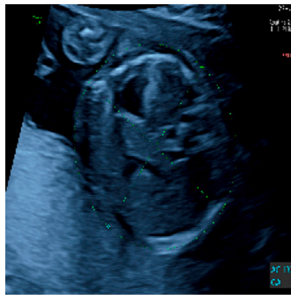
**GL:** measuring around the baby's belly and the stomach umbilical vein rib rib and that is the baby's belly

**RG:** this is the baby's belly and stomach and let us us measure around the baby's belly and the baby's belly has good size

**PG:** this is the baby's belly and the stomach and let us measure around the baby's belly and the baby's belly has good size

**UL:** this is the head

(a) A fetal ultrasound frame of an abdomen with corresponding captions.



**GT:** this is where we measure the abdominal circumference

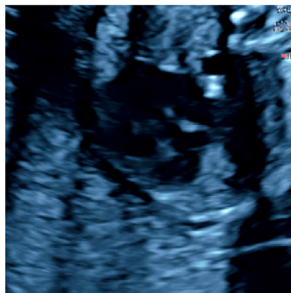
**GL:** is belly and stomach and umbilical vein rib rib and spine and spine and spine

**RG:** stomach and this is the baby's belly and the stomach and let us measure around the baby's belly and the baby's belly has good size

**PG:** this is the baby's belly and the stomach and let us measure around the baby's belly and the baby's belly has good size

**UL:** this is the head

(b) A fetal ultrasound frame of an abdomen with corresponding captions. This frame is a zoomed in version of the one in (a).



**GT:** we can see the right and left ventricles and nice contraction and this is the right ventricle outflow tract and so we can see normal pulmonary valve

**GL:** you can see the valves here and there and this this is the tricuspidal valve and you can see the tricuspidal valve opening and and opening and opening and closing and good tricuspidal valve

**RG:** the left ventricular outflow tract and right right ventricular outflow tract is there good good

**PG:** left ventricular outflow tract and the right ventricular outflow tract are looking very good

**UL:** this is the head

(c) A fetal ultrasound frame of a heart with corresponding captions

**Fig. 7.** Qualitative examples of captions generated by the three different model configurations with varying levels of gaze involvement are shown. The frame shown is the central frame in the video clip being captioned. 'GT' stands for 'Ground Truth'. This caption comes from the recorded speech of the sonographer. 'GL' stands for 'Gaze-Less'. This caption was generated by a model that does not use eye tracking information. 'RG' stands for 'Real Gaze'. This caption was generated by a model that uses the real eye tracking data in the captioning process in the form of real gaze attention maps. 'PG' stands for 'Predicted Gaze'. 'UL' is for 'UniVL'. This caption was generated by a model that uses predicted gaze saliency maps in the captioning process. Words are underlined to show where generated captions exactly match the ground truth. See Tables 5–7 for the quantitative results of these generated captions.

## A.2. Fine-tuning the CNN

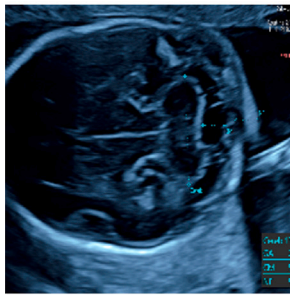
Anatomical class labels were used to fine-tune the VGG16 (Simonyan and Zisserman, 2014) responsible for extracting frame features. These labels were obtained through manual labelling of video clips in the US scan videos based on the viewed anatomies as described in Sharma et al. (2019).

## A.3. Additional vocabulary information

The training vocabulary had 330 unique words. 261 of those would not be considered stop words by the natural language toolkit (NLTK) (Bird et al., 2009). 'Baby' is the most common non-stop word repeated

71 times (2.3% of total text). The 2nd most common non-stop word is 'head' with 47 occurrences (1.5% of total text). The least frequent non-stop words include 'coronally', 'top', 'sort', 'put', 'bits', 'lines', 'tend', 'vulva', 'arms', 'time', 'lying', 'internal', 'lower', 'fertility', 'index', 'add', 'views', 'bound', 'goes', 'spectrum', 'vermis', 'thalamus', 'choroid', 'plexus', 'sized', 'symmetrical', 'close', 'aortic', 'hands', 'ultrasound', 'signal', 'profile', 'connects', 'parts', 'able', 'separately', 'find', 'precisely', 'appreciate', 'wish', 'line', 'iliac', 'sign', 'still', and 'get' with only one occurrence for each of them. 'Go', 'open', 'shaped', 'white', 'know', 'straight', 'alright', 'space', and 'midline' are among the words with a median number of occurrences (2). The text dataset has a lexical diversity of 0.13. This number indicates that 13% of the all the text consist of unique words (Bird et al., 2009). Collocations are words





**GT:** this is this cisterna magna also looks good size and nuchal fold yeah measurements are good

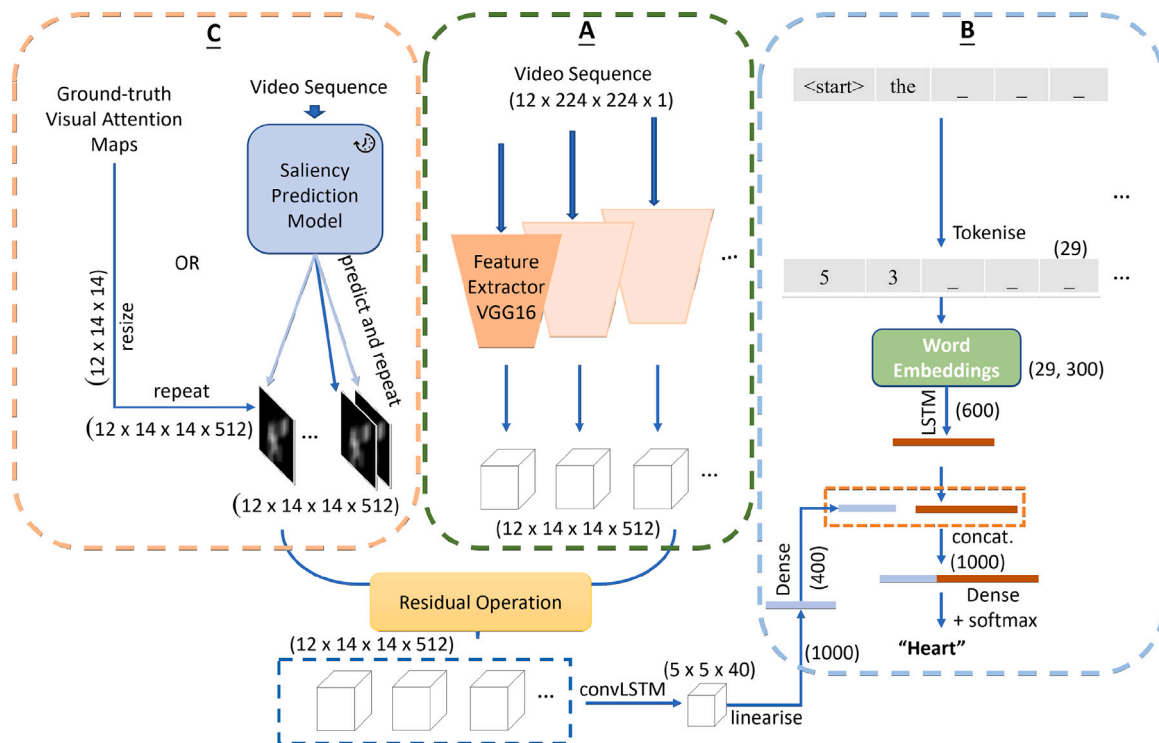
**GL:** baby's head and brain side to side and around baby's head and this is the

**RG:** we can see baby's skull we can see the midline of the brain and this is where we can see the cavum septum pellucidum and now the posterior lateral ventricle and now measuring baby's head side to side and around baby's head and this is the posterior lateral ventricle normal size

**PG:** this is baby's head and brain and measuring the head side to this is the cisterna magna and nuchal fold look good

(a) A fetal ultrasound frame of a head with corresponding captions.

**Fig. 8.** Another qualitative example of captions generated by the three different model configurations with varying levels of gaze involvement are shown. The frame shown is the central frame in the video clip being captioned. 'GT' stands for 'Ground Truth'. This caption comes from the recorded speech of the sonographer. 'GL' stands for 'Gaze-Less'. This caption was generated by a model that does not use eye tracking information. 'RG' stands for 'Real Gaze'. This caption was generated by a model that uses the real eye tracking data in the captioning process in the form of real gaze attention maps. 'PG' stands for 'Predicted Gaze'. This caption was generated by a model that uses predicted gaze saliency maps in the captioning process. Words are underlined to show where generated captions exactly match the ground truth.



**Fig. 9.** The architecture of the multi-modal model is shown in the this figure in more detail. The specifics of the residual operation that is performed on the outputs of Blocks A and C are shown on the right side of this figure. The gazeless model configuration only includes Blocks A and B. Block A represents the branch of the model where the spatial feature information is extracted from the video clip for each of its sampled frames by a VGG16 CNN. Block B represents the branch handling text information. The sequence of words generated so far are tokenized and embedded with a Word2vec embedding vector before being passed as input to an LSTM-RNN. The last hidden state of this LSTM-RNN is concatenated with the linearised feature vector from the convolutional LSTM. The real gaze and predicted gaze model configurations also include Block C. In Block C, either the ground truth visual attention maps are used (in the case of the real gaze configuration) or predicted gaze saliency maps are predicted and then used for each sampled frame in a video clip by a previously trained saliency prediction model. The extracted feature blocks from Block A and the gaze maps from Block C are combined together through the residual operation shown in detail on the right side of the figure. In the gazeless configuration, there is no residual operation to be performed on the sequence of feature blocks. They are passed directly to the convolutional LSTM.

that regularly occur together in the text. In our dataset, these include expected phrases such as 'cisterna magna', 'nuchal fold', 'outflow tract', 'three vessel', 'lateral ventricle', 'abdominal circumference', 'four chamber', 'trachea view', 'umbilical vein' as well as unexpected phrases such as 'looks good' and 'baby's belly'. See Figs. 5 and 6 for more information on the occurrence of common words.

## References

- Allaoui, I., Ben Ahmed, M., Benamrou, B., Ouardouz, M., 2018. Automatic caption generation for medical images. In: Proceedings of the 3rd International Conference on Smart City Applications. pp. 1–6.
- Alsharid, M., El-Bouri, R., Sharma, H., Drukker, L., Papageorgiou, A.T., Noble, J.A., 2020. A curriculum learning based approach to captioning ultrasound images.

- In: Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Springer, pp. 75–84.
- Alsharid, M., El-Bouri, R., Sharma, H., Drukker, L., Papageorgiou, A.T., Noble, J.A., 2021. A course-focused dual curriculum for image captioning. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 716–720.
- Alsharid, M., Sharma, H., Drukker, L., Chatelain, P., Papageorgiou, A.T., Noble, J.A., 2019. Captioning ultrasound images automatically. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 338–346.
- Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B., 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artificial Intelligence Res.* 55, 409–442.
- Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc..
- Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Cai, Y., 2019. Deep Learning Sonographer Visual Attention (Ph.D. thesis). University of Oxford.
- Cai, Y., Droste, R., Sharma, H., Chatelain, P., Drukker, L., Papageorgiou, A.T., Noble, J.A., 2020. Spatio-temporal visual attention modelling of standard biometry plane-finding navigation. *Med. Image Anal.* 65, 101762.
- Cai, Y., et al., 2018a. Multi-task SonoEyeNet: detection of fetal standardized planes assisted by generated sonographer attention maps. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 871–879.
- Cai, Y., et al., 2018b. Sonoeyenet: Standardized fetal ultrasound plane detection informed by eye tracking. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 1475–1478.
- Cuturi, M., Blondel, M., 2017. Soft-DTW: a differentiable loss function for time-series. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 894–903.
- Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al., 2018. Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 720–736.
- Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D., 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Comput. Vis. Image Underst.* 163, 90–100.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Droste, R., Cai, Y., Sharma, H., Chatelain, P., Papageorgiou, A.T., Noble, J.A., 2019. Towards capturing sonographic experience: cognition-inspired ultrasound video saliency prediction. In: Annual Conference on Medical Image Understanding and Analysis. Springer, pp. 174–186.
- Drukker, L., Droste, R., Chatelain, P., Noble, J., Papageorgiou, A., 2020a. Expected-value bias in routine third-trimester growth scans. *Ultrasound Obstet. Gynecol.* 55 (3), 375–382.
- Drukker, L., Droste, R., Chatelain, P., Noble, J.A., Papageorgiou, A.T., 2020b. Safety indices of ultrasound: adherence to recommendations and awareness during routine obstetric ultrasound scanning. *Ultraschall Der Medizin-European J. Ultrasound* 41 (02), 138–145.
- Drukker, L., Droste, R., Noble, A., Papageorgiou, A., 2020c. VP40. 20: Standard biometric planes: what are the salient anatomical landmarks? *Ultrasound Obstet. Gynecol.* 56, 235.
- Drukker, L., Sharma, H., Droste, R., Alsharid, M., Chatelain, P., Noble, J.A., Papageorgiou, A.T., 2021. Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. *Sci. Rep.* 11 (1), 1–12.
- Elliott, D., Keller, F., 2013. Image description using visual dependency representations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1292–1302.
- Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D., 2010. Every picture tells a story: Generating sentences from images. In: European Conference on Computer Vision. Springer, pp. 15–29.
- Goldberg, Y., 2016. A primer on neural network models for natural language processing. *J. Artificial Intelligence Res.* 57, 345–420.
- Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep Learning. Vol. 1, (2), MIT press Cambridge.
- Google Cloud, 2019. Google cloud speech-to-text - speech recognition. URL <https://cloud.google.com/speech-to-text/>.
- Guadarrama, S., Krishnamoorthy, N., Malkarnkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K., 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2712–2719.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
- Kingma, D., 2015. Ba J. Adam: a method for stochastic optimization. In: The International Conference on Learning Representations.
- Kirwan, D., 2010. NHS fetal anomaly screening programme. National Stand. Guid. Engl. 18.
- Kojima, A., Tamura, T., Fukunaga, K., 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *Int. J. Comput. Vis.* 50 (2), 171–184.
- Kougia, V., Pavlopoulos, J., Androutsopoulos, I., 2019. A survey on biomedical image captioning. *arXiv preprint arXiv:1905.13302*.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36 (4), 1234–1240.
- Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.
- Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M., 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.
- Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., Sivic, J., 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640.
- Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119.
- Naseem, U., Khushi, M., Reddy, V., Rajendran, S., Razzak, I., Kim, J., 2021. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In: 2021 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–7.
- NHS, 2018. NHS fetal anomaly screening programme handbook 2018. URL [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/749742/NHS\\_fetal\\_anomaly\\_screening\\_programme\\_handbook\\_FINAL1.2.18.10.18.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/749742/NHS_fetal_anomaly_screening_programme_handbook_FINAL1.2.18.10.18.pdf).
- Ordonez, V., Kulkarni, G., Berg, T.L., 2011. Im2text: Describing images using 1 million captioned photographs. In: Advances in Neural Information Processing Systems. pp. 1143–1151.
- Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y., 2016a. Jointly modeling embedding and translation to bridge video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4594–4602.
- Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y., 2016b. Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1029–1038.
- Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 311–318.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J., 2010. Collecting image annotations using amazon's mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. pp. 139–147.
- Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B., 2013. Translating video content to natural language descriptions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 433–440.
- Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., Schiele, B., 2014. Coherent multi-sentence video description with variable level of detail. In: German Conference on Pattern Recognition. Springer, pp. 184–195.
- Saab, K., Hooper, S.M., Sohoni, N.S., Parmar, J., Pogatchnik, B., Wu, S., Dunnmon, J.A., Zhang, H.R., Rubin, D., Ré, C., 2021. Observational supervision for medical image classification using gaze data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 603–614.
- Sharma, H., Droste, R., Chatelain, P., Drukker, L., Papageorgiou, A., Noble, J., 2019. Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 987–990.
- Sharma, H., Drukker, L., Chatelain, P., Droste, R., Papageorgiou, A.T., Noble, J.A., 2021. Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos. *Med. Image Anal.* 69, 101973.

- Siersdorfer, S., San Pedro, J., Sanderson, M., 2009. Automatic video tagging using content redundancy. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 395–402.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sugano, Y., Bulling, A., 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*.
- Tanti, M., Gatt, A., Camilleri, K.P., 2017. What is the role of recurrent neural networks (RNNs) in an image caption generator? *arXiv preprint arXiv:1708.02043*.
- Tanti, M., Gatt, A., Camilleri, K.P., 2018. Where to put the image in an image caption generator. *Nat. Lang. Eng.* 24 (3), 467–489.
- Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3156–3164.
- Wu, Z., Yao, T., Fu, Y., Jiang, Y.-G., 2017. Deep learning for video classification and captioning. In: *Frontiers of Multimedia Research*. pp. 3–29.
- Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: *Advances in Neural Information Processing Systems*. pp. 802–810.
- Xiong, Y., Du, B., Yan, P., 2019. Reinforced transformer for medical image captioning. In: *International Workshop on Machine Learning in Medical Imaging*. Springer, pp. 673–680.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015a. Show, attend and tell: Neural image caption generation with visual attention. In: *International Conference on Machine Learning*. pp. 2048–2057.
- Xu, J., Mei, T., Yao, T., Rui, Y., 2016. MSR-VTT: A large video description dataset for bridging video and language. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5288–5296.
- Xu, R., Xiong, C., Chen, W., Corso, J.J., 2015b. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: *AAAI*. Vol. 5, Citeseer, p. 6.
- Yao, T., Mei, T., Ngo, C.-W., Li, S., 2013. Annotation for free: Video tagging by mining user search behavior. In: *Proceedings of the 21st ACM International Conference on Multimedia*. pp. 977–986.
- You, Q., Jin, H., Wang, Z., Fang, C., Luo, J., 2016. Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4651–4659.
- Young, P., Lai, A., Hodosh, M., Hockenmaier, J., 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* 2, 67–78.
- Yu, Y., Choi, J., Kim, Y., Yoo, K., Lee, S.-H., Kim, G., 2017. Supervising neural attention models for video captioning by human gaze data. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 490–498.
- Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W., 2016. Video paragraph captioning using hierarchical recurrent neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4584–4593.
- Zeng, X.-H., Liu, B.-G., Zhou, M., 2018. Understanding and generating ultrasound image description. *J. Comput. Sci. Tech.* 33 (5), 1086–1100.
- Zeng, X., Wen, L., Liu, B., Qi, X., 2019. Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing*.
- Zhang, Y., et al., 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* 6 (1), 1–9.
- Zhao, B., Li, X., Lu, X., 2019. CAM-RNN: Co-attention model based RNN for video captioning. *IEEE Trans. Image Process.* 28 (11), 5552–5565.
- Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J., 2020. Unified vision-language pre-training for image captioning and vqa. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, (07), pp. 13041–13049.
- Zhu, X., Li, L., Liu, J., Peng, H., Niu, X., 2018. Captioning transformer with stacked attention modules. *Appl. Sci.* 8 (5), 739.